This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

PALM: Progress-Aware Policy Learning via Affordance Reasoning for Long-Horizon Robotic Manipulation

Yuanzhe Liu[1,2] Jingyuan Zhu[1] Yuchen Mo[2] Gen Li[3] Xu Cao[2] Jin Jin[4] Yifan Shen[2]
Zhengyuan Li[2] Tianjiao Yu[2] Wenzhen Yuan[2] Fangqiang Ding[5] Ismini Lourentzou[2]

[1]University of Pennsylvania [2]University of Illinois Urbana-Champaign
[3]Nanyang Technological University [4]University of Oxford [5]Massachusetts Institute of Technology

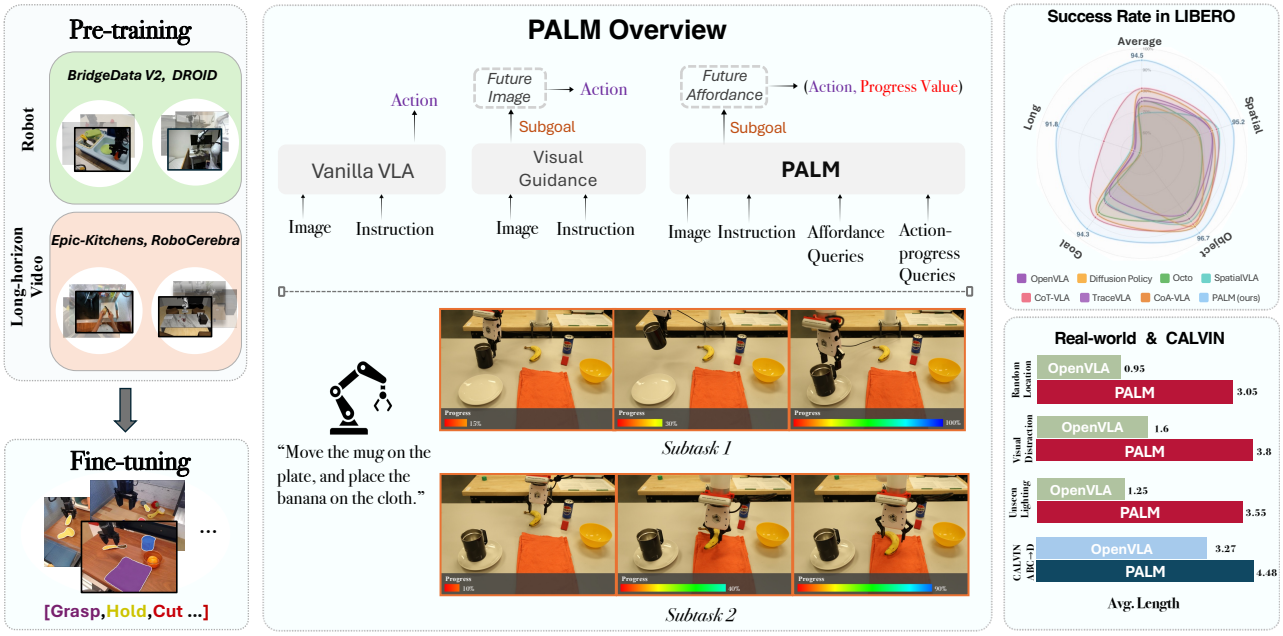


Figure 1. In contrast to vanilla VLAs that directly map inputs to actions or to predictive methods that forecast dense future images, **PALM** introduces learnable queries to forecast a structured set of future affordances. Conditioned on these affordances, a diffusion-based policy jointly decodes the robot's action and a continuous progress value, enabling temporal state tracking and seamless subtask transitions in multi-step, long-horizon tasks. Our training strategy utilizes large-scale pre-training on both robot datasets (BridgeDataV2 [113], DROID [54]) and long-horizon video data (EPIC-KITCHENS [20], RoboCerebra [33]), followed by fine-tuning on our collected human-annotated affordance dataset. Consequently, **PALM** attains state-of-the-art performance on long-horizon simulation benchmarks (CALVIN ABC→D [88], LIBERO-LONG [77]) and demonstrates strong results under real-world long-horizon generalization settings.

Abstract

Recent advancements in vision-language–action (VLA) models have shown promise in robotic manipulation, yet they continue to struggle with long-horizon, multi-step tasks. Existing methods lack internal reasoning mechanisms that can identify task-relevant interaction cues or track progress within a subtask, leading to critical execution errors such as repeated actions, missed steps, and premature termination. To address these challenges, we introduce **PALM**, *a VLA framework that structures policy learning around interaction-centric affordance reasoning and subtask progress cues.* PALM *distills complementary affordance representations that capture object relevance, contact geometry, spatial placements, and motion dynamics, and serve as task-relevant anchors for visuomotor control. To further stabilize long-horizon execution,* PALM *predicts continuous within-subtask progress, enabling seamless subtask transitions. Across extensive simulation and real-world experiments,* PALM *consistently outperforms baselines, achieving a 91.8% success rate on LIBERO-LONG, a 12.5% improvement in average length on CALVIN ABC→D, and a 2× improvement over real-world baselines across three long-horizon generalization settings.*
PLAN Lab https://plan-lab.github.io/palm

1. Introduction

Robot learning has made significant strides in developing generalizable policies for diverse tasks and environments [5, 12, 23, 25, 28, 35, 72, 117, 118, 125, 136, 144]. Much of this progress is driven by Vision-Language-Action (VLA) models, which leverage pre-trained vision-language backbones to map visual observations and language instructions directly to robot actions [7, 8, 10, 19, 31, 45, 55, 67, 69, 92, 103, 111, 121, 138, 150, 155]. However, current VLA methods are fundamentally limited to short-horizon manipulation and struggle with long-horizon, multi-step planning in dynamic scenes. For example, on "clean a cluttered table," state-of-the-art policies typically succeed initially but fail mid-task, unable to reliably complete the full sequence.

A fundamental limitation is the absence of structured affordance cues [39, 42, 58, 107, 135] and explicit state tracking [16, 50]. Although existing models may infer the final goal and produce intermediate actions [18, 38, 112, 143, 146, 148], they lack internal representations that disambiguate which object should be targeted next, which part or region is relevant for interaction, where items should be placed or moved, or what motion is appropriate for the upcoming step. Consequently, many visually similar states become ambiguous, obscuring the underlying task stage and destabilizing long-horizon control. In addition, existing VLAs lack mechanisms for continuously estimating progress within a subtask. Without a persistent online notion of advancement, the policy cannot reliably decide whether to continue, switch stages, or terminate. This absence of temporal grounding leads to characteristic long-horizon failure modes: repeated or unnecessary actions, skipped required subtasks, premature termination, and even declaring success in incorrect states.

To address these gaps, we introduce **PALM**, a novel end-to-end framework for learning scalable, long-horizon manipulation. As illustrated in Fig. 1, PALM integrates perception, action, and progress within a closed loop and is built around two complementary capabilities. First, PALM predicts structured future affordances encoded as latent queries, which are refined through a block-wise structured attention mechanism and specialize into four types, capturing object relevance (global), contact-level geometry (local), candidate placement regions (spatial), and plausible motion trajectories for the next interaction step (dynamic). These latents form a compact and task-relevant representation of the evolving scene state. Second, PALM estimates fine-grained subtask progress through a continuous progress signal. Conditioned on both the affordance latents and the current multimodal context, the action policy employs a diffusion transformer to jointly predict action and progress sequences. By combining structured affordance prediction with progress-aware action generation, PALM maintains stable, coherent behavior across long, multi-step manipulation sequences.

We conduct large-scale pre-training on both robot [54, 113] and long-horizon video data [20, 33], and then fine-tune the model on human-annotated robot trajectories. Through extensive experiments in both simulation and real-world settings, PALM demonstrates clear and consistent gains in long-horizon manipulation. On two widely used long-horizon simulation benchmarks, it improves the success rate on CALVIN ABC→D [88] by 12.5% over prior state-of-the-art baselines and reaches 91.8% success on LIBERO-LONG [77]. Beyond simulation, we design three types of challenging real-world generalization tests for long-horizon scenarios that vary in localization, lighting, and visual distractors. Under all conditions, PALM consistently demonstrates strong performance and robustness with limited fine-tuning data. Our contributions are as follows:

- We introduce **PALM**, a unified VLA framework that integrates structured affordance reasoning and progress-aware policy generation to enable reliable execution across long-horizon, contact-rich manipulation tasks.
- We propose two novel complementary modules: (1) a fine-grained affordance predictor that acts as an intermediate implicit reasoning step, producing structured and task-relevant representations, and (2) a progress-aware inverse-dynamics module that couples action generation with subtask progress estimation, ensuring temporally coherent execution over long action sequences.
- We conduct comprehensive evaluations both in simulation and the real world, achieving a 12.5% improvement on CALVIN ABC→D over prior state-of-the-art, 91.8% success on LIBERO-LONG, and consistent gains across three real-world long-horizon generalization tests.

2. Related Work

Vision-Language-Action Models. Several works repurpose pre-trained Vision-Language Models (VLMs) [1, 22, 41, 63, 71, 90, 104] into VLA policies that map visual observations and language instructions to low-level robot actions [6, 21, 23, 47, 52, 74, 75, 105, 123, 149] by fine-tuning on large-scale robotics datasets [27, 54, 94]. A prominent paradigm, pioneered by the RT series [4, 10, 155], formulates action generation as autoregressive prediction over tokenized sequences [9, 55, 70, 154]. In parallel, diffusion-based action generators [19, 37, 53, 80, 137] treat control as a denoising process in continuous trajectory spaces, producing smoother temporal dynamics. Despite their success, both paradigms rely on direct action prediction, which lacks explicit reasoning and fine-grained representations of spatial or physical dynamics. Subsequent works address these limitations by incorporating future prediction through goal-image generation [26, 31, 38, 117, 136, 150] or integrated forecasting [112, 126, 143, 148, 153], or by enhancing spatio-temporal grounding via keypoint prediction [42, 135] and historical visual traces [93]. In contrast, our work introduces a closed perception–action–progress loop that improves long-

horizon manipulation by integrating affordance and subtask progress reasoning into the VLA.

Imitation Learning with Progress Supervision. Early imitation learning approaches for long-horizon tasks relied on explicit task decomposition, such as symbolic planning [29, 48] or predefined visual keyframes [11, 97]. Recent works have shifted towards learning continuous progress representations [3, 73, 108, 139, 152] directly from large-scale video data [85, 86, 91]. These learned progress metrics can be used to reweight behavior cloning [17, 106, 119], assigning higher importance to demonstrations that yield greater progress. Compared to computationally expensive bootstrapping methods [124] or strategies that rely on a small, high-quality subset for data labeling [114, 120], this intrinsic progress weighting offers improved scalability and stability. Our approach augments the action space with a scalar progress indicator, used to supervise policy learning for improved long-horizon manipulation consistency.

Affordance Representations for Robotic Manipulation. Representing affordances is a core problem in robot learning [34, 92, 109, 110]. Early work generated structured visual outputs from sensor data, predicting affordances as heatmaps over interaction regions [2, 15, 30, 64, 65, 76, 89, 130] or as bounding boxes and keypoints to localize relevant parts [40, 66, 78, 87, 99]. With VLMs, affordance reasoning became open-vocabulary, either via retrieval from knowledge bases [49, 58] or via fine-tuning on affordance datasets [83, 84, 98]. Recent VLA models further integrate affordance into the policy loop by injecting visual interaction traces [151], predicting affordance plans [92], or encoding affordance types into reasoning chains [67]. Our approach predicts multiple affordance types as structured latent tokens that provide implicit task-aware representations for progress estimation and subtask policy learning.

3. Method

3.1. Problem Formulation

Long-horizon, language-conditioned manipulation requires policies that remain temporally consistent across sub-tasks with heterogeneous dynamics. Let a robot operate over a distribution of tasks $\mathcal{T} = \{\tau_k\}_{k=1}^{K}$. Each task $\tau \in \mathcal{T}$ defines an observation–action distribution $p\left(o_t, a_t \mid \tau\right)$ and an implicit temporal phase progression. Differences in background, lighting, object configuration, and viewpoint, compounded by cumulative control errors, alter the observation space. As a result, when sub-policies are trained in isolation, the terminal state of one often deviates from the expected start state of the next, leading to unstable transitions. Conventional behavior cloning that directly fits $\pi\left(a_t \mid o_t\right)$ on a mixture of demonstrations collapses these distinct phases, causing repetition or skipped subtasks during rollout. As illustrated in Fig. 2, PALM addresses long-horizon stability through two tightly coupled components: ① **a fine-grained affordance prediction module** that anticipates future interaction cues by estimating an affordance-centric latent at future offsets $t+n$, $f_{\text{aff}} : \mathcal{O} \times \mathcal{T} \to \mathcal{F}$, serving as a task-relevant intermediate representation that stabilizes perception (§3.3), and ② **a progress-aware policy** $\pi : \mathcal{O} \times \mathcal{T} \to \tilde{\mathcal{A}}$, $\tilde{\mathcal{A}} = \mathcal{A} \times \mathcal{P}$, where $\mathcal{P} \subset [0, 1]$ is a continuous progress space. At time t, given observations $o_t \in \mathcal{O}$, and task specification $\tau \in \mathcal{T}$, and conditioned on the predicted affordance latent, the policy jointly decodes an action $a_t \in \mathcal{A}$ alongside a scalar $p_t \in \mathcal{P}$ that encodes progress within the current subtask and serves as a temporal regularizer that mitigates phase ambiguity and enables smooth subtask transitions without relying on separate planners or hierarchical controllers (§3.4).

3.2. PALM Architecture

Multi-Modal Encoders. PALM processes three synchronized inputs: a language instruction l, an image observation o_t, and a robot state s_t. Instructions are embedded using a CLIP text encoder [102], observations are encoded with a Masked Autoencoder [36] and downsampled by a Perceiver Resampler [46] to retain task-relevant visual tokens. The robot state is projected through a lightweight MLP. The resulting tokens are concatenated into a single sequence with time index t to form a unified multimodal sequence passed to the transformer backbone.

Backbone and Learnable Queries. The backbone follows a GPT-2-style transformer [101] to integrate the token sequence with causal and cross-modal attention. On top of this backbone, PALM introduces two learnable query sets that instantiate the affordance and progress-aware policy:

◇ ***Fine-grained affordance queries*** comprise four subqueries `<Global>`, `<Local>`, `<Spatial>`, and `<Dynamic>`, that attend over language, vision, and state tokens to extract task-relevant representation at complementary scales and produce the affordance-centric latent

$$\hat{\mathbf{F}}_{t+n} = f_{\text{aff}}(l, o_t, s_t) \in \mathcal{F}. \tag{1}$$

◇ ***Action–progress queries*** pool control-relevant context and condition on $\hat{\mathbf{F}}_{t+n}$ to support progress-aware inverse-dynamics decoding for temporally consistent long-horizon control. Building on prior inverse-dynamics formulations [18, 38, 112], these queries aggregate current observations with the predicted affordance latent to infer action sequences that align with the forecasted interaction state.

Decoders. The action head is a denoising diffusion transformer [96] that conditions on the action–progress queries and the affordance latent to generate a horizon-n trajectory:

$$\left(\hat{a}_{t:t+n-1}, \hat{p}_{t:t+n-1}\right) = \text{DiT}\left(l, o_t, s_t, \hat{\mathbf{F}}_{t+n}\right), \tag{2}$$

where $\hat{a}_{t:t+n-1}$ is the predicted action sequence and $\hat{p}_{t:t+n-1}$ the predicted sequence of scalar progress values at each step. Details can be found in Appendix A.

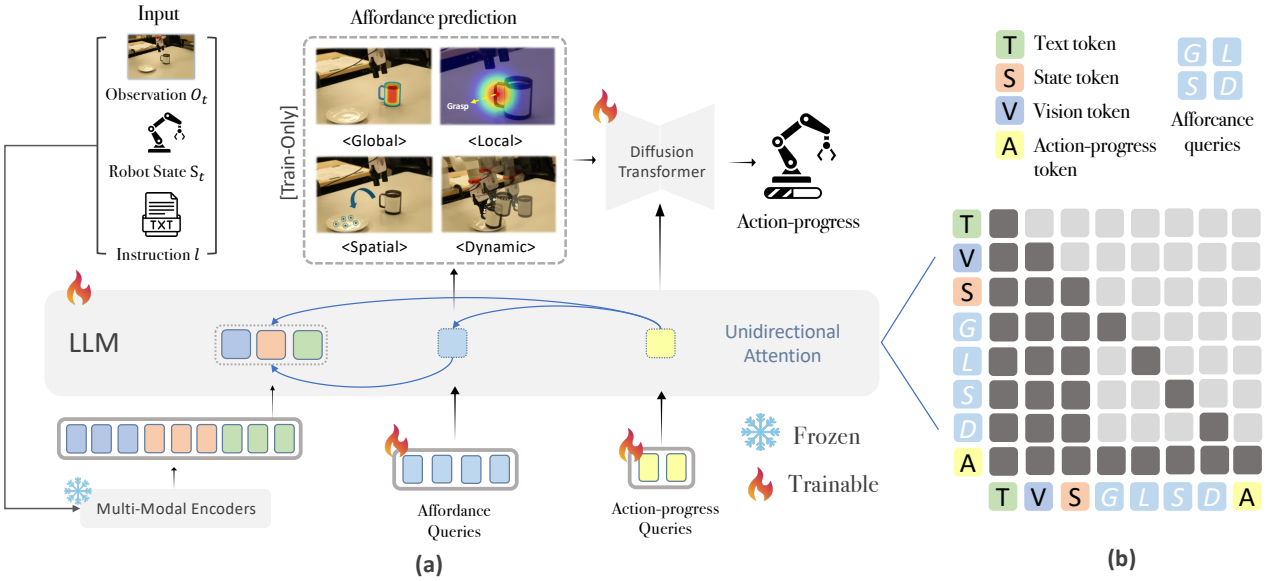


Figure 2. **PALM Overview. (a) Model Architecture:** Given a language instruction l, observation o_t, and robot state s_t, PALM encodes each modality using frozen encoders to obtain text, visual, and state tokens. These tokens are fused by a GPT-style transformer with *unidirectional attention* and two specialized query sets: *fine-grained affordance* and *action–progress*. During training, affordance queries attend to context tokens to predict foresight $\hat{\mathbf{F}}_{t+n}$ with four supervised heads (<Global>, <Local>, <Spatial>, <Dynamic>) that ground future scene understanding. At inference, the affordance heads are removed; the action–progress query attends to both context and affordance foresight to condition a diffusion transformer that predicts action $\hat{a}_{t:t+n-1}$ and progress $\hat{p}_{t:t+n-1}$ trajectories for continuous control. **(b) Structured Attention:** Affordance subqueries attend only to shared context tokens to stay disentangled, while both query types use causal attention to preserve temporal consistency.

3.3. Fine-Grained Affordance Prediction

Global Prediction. Global affordance provides a high-level semantic prior that identifies *which* object in the scene is relevant to the instruction and *where* it is located. Given a command such as "pick up the cup," the model parses the referent (*i.e.*, the instruction-specified object) at time t and predicts, for a future offset $t+n$, an object-centric prior that marks the referred instance and its approximate region. This global prior anchors all subsequent affordance cues and constrains local reasoning to the correct region of interest.

Targets are obtained by first resolving the referent via Grounding DINO [81] and then segmenting the instance with SAM [57] to obtain a binary mask on the observation lattice. An object representation is further extracted by masked pooling over a frozen image encoder within this region. During training, the predicted future mask and object feature are supervised against these targets. The training objective for global affordance tokens is then formulated as:

$$\mathcal{L}_{\text{global}} = \mathcal{L}_{\text{FL}}\left(\mathcal{M}^{\text{global}}_{t+n}, \hat{\mathcal{M}}^{\text{global}}_{t+n}\right) + \mathcal{L}_{\text{Dice}}\left(\mathcal{M}^{\text{global}}_{t+n}, \hat{\mathcal{M}}^{\text{global}}_{t+n}\right) \tag{3}$$

where $\hat{\mathcal{M}}^{\text{global}}_{t+n} = f_{\text{global}}\left(l, o_t, s_t\right)$ denotes the predicted global affordance mask at time $t+n$, and $\mathcal{M}^{\text{global}}_{t+n}$ is the corresponding binary target mask on the image (observation) lattice. Here, $\mathcal{L}_{\text{FL}}$ denotes the pixel-wise focal loss computed over image domain Ω, and $\mathcal{L}_{\text{Dice}}$ denotes the soft Dice loss.

Local Prediction. Local affordance captures fine-grained geometric localization affordance cues necessary for precise contact reasoning. Conditioned on the region highlighted by the global prior, it analyzes high-frequency local visual structure, such as edges, textures, and part geometry, to predict, at time $t+n$, a dense contact-likelihood distribution over the object of interest. Following GLOVER++[84], annotated contact points are converted into Gaussian heatmaps that serve as continuous supervision targets. The model is trained to match this target heatmap distribution. The training objective for local affordance tokens is thus formulated as:

$$\mathcal{L}_{\text{local}} = \mathcal{L}_{\text{FL}}\left(\mathcal{M}^{\text{local}}_{t+n}, \hat{\mathcal{M}}^{\text{local}}_{t+n}\right) + \mathcal{L}_{\text{KL}}\left(\tilde{\mathcal{M}}^{\text{local}}_{t+n}, \tilde{\hat{\mathcal{M}}}^{\text{local}}_{t+n}\right), \tag{4}$$

$$\tilde{\mathcal{M}} = \frac{\mathcal{M}}{\sum_{(i,j)\in\Omega} \mathcal{M}^{(i,j)} + \varepsilon}, \tag{5}$$

where $\hat{\mathcal{M}}^{\text{local}}_{t+n} = f_{\text{local}}\left(l, o_t, s_t\right)$ is the predicted future local contact-likelihood map at $t+n$, $\mathcal{M}^{\text{local}}_{t+n}$ is the soft Gaussian target heatmap, and $\tilde{\mathcal{M}}$ denotes the ℓ_1-normalized map used in the KL term with a small $\varepsilon > 0$.

Spatial Prediction. Spatial affordance converts underspecified spatial language into executable placement proposals

that remain robust under layout changes. Instead of memorizing a single coordinate, the model predicts a small set of candidate placement points at time $t+n$, each representing a plausible region for placing the object. Ground-truth targets are obtained by first converting the instruction into spatial semantics using SpatialVLM [13] and then sampling executable 2D coordinates with RoboPoint [135]. The training objective minimizes a set-matching objective that aligns each target point with its closest predicted candidate:

$$\mathcal{L}_{\text{spatial}} = \frac{1}{C_{t+n}} \sum_{c=1}^{C_{t+n}} \min_{1 \le m \le M} \left\| \hat{\mathbf{p}}_{t+n}^{(m)} - \mathbf{p}_{t+n}^{(c)} \right\|_2^2 \tag{6}$$

where $\hat{\mathcal{S}}_{t+n} = \{\hat{\mathbf{p}}_{t+n}^{(m)}\}_{m=1}^{M}$ and $\mathcal{S}_{t+n} = \{\mathbf{p}_{t+n}^{(c)}\}_{c=1}^{C_{t+n}}$ are the set of M predicted normalized 2D candidates and C_{t+n} target placement points at $t+n$, respectively.

Dynamic Prediction. Dynamic affordance identifies the pixels corresponding to the robot gripper and movable objects, and predicts how these regions will evolve over time. Its goal is to predict and establish the statistical associations among the current scene, the language instruction, and the actions required to realize the predicted motion. To construct supervision, we apply a grid-based tracking protocol: an $N \times N$ visual grid of query points is placed on the first frame of a short history at $t-\delta$. CoTracker [51] follows each point forward in time. We compute cumulative displacement for each trajectory and retain those that exceed a threshold, ensuring that the dynamic mask captures genuine motion rather than static background or tracker jitter. The selected trajectories are then rasterized at $t+n$ into a dynamic region. Given inputs at time t, the model predicts a future dynamic probability map that highlights pixels likely to belong to the gripper or other moving objects at $t+n$. Training encourages calibrated per-pixel probabilities and alignment with the tracked motion regions. The training objective for dynamic affordance tokens is formulated as a masked reconstruction loss using a latent-variable model:

$$\begin{aligned} \mathcal{L}_{\text{dynamic}} = \mathbb{E}_{\mathbf{z} \sim Q_\phi(\mathbf{z} \mid x_{t+n}^{\mathcal{M}})} \Big[-\log P_\psi \Big(x_{t+n}^{\mathcal{M}} \mid \mathbf{z} \Big) \Big] \\ + \beta \, \mathrm{KL} \Big(Q_\phi \Big(\mathbf{z} \mid x_{t+n}^{\mathcal{M}} \Big) \,\|\, p(\mathbf{z}) \Big) \end{aligned} \tag{7}$$

where $\mathcal{M}_{t+n}$ is the dynamic region mask from tracking, $x_{t+n}^{\mathcal{M}}$ is the future frame restricted to those masked pixels, $Q_\phi(\mathbf{z} \mid x_{t+n}^{\mathcal{M}})$ is the posterior, $P_\psi(x_{t+n}^{\mathcal{M}} \mid \mathbf{z})$ the decoder likelihood, $p(\mathbf{z})$ the prior, and β the KL weight.

3.4. Progress-aware Policy via Inverse Dynamics

In addition to predicting where to act via affordances, we introduce a progress-aware prediction task that estimates how far execution has advanced within the current subtask. At each time step, PALM first infers the active subtask stage from the affordance latent and derives a stage embedding. Conditioned on this embedding, the model predicts a scalar $p_t \in [0, 1]$ that quantifies within-stage completion. We append this scalar to the action output so the policy jointly predicts (a_t, p_t) under a shared multimodal context. This explicit progress signal reduces ambiguity in long-horizon control: visually similar observations may correspond to different actions depending on stage, and p_t disambiguates these cases by providing a continuous indicator of "where we are", stabilizing learning and execution by encouraging monotonic, stage-consistent evolution of the latent state and by smoothing transitions at sub-policy boundaries without relying on separate high-level controllers.

Classical inverse dynamics predicts the action $\hat{a}_t$ bridging two temporally ordered observations (o_t, o_{t+1}). We extend this to predict an n-step action–progress sequence conditioned on the current inputs and a single-step affordance latent. We instantiate f_{inv} as a denoising diffusion transformer that conditions on the current observation o_t, the instruction l, the robot state s_t, and the predicted affordance latent $\hat{\mathbf{F}}_{t+n} = f_{\text{aff}}(l, o_t, s_t)$ to generate

$$(\hat{a}_{t:t+n-1}, \hat{p}_{t:t+n-1}) = f_{\text{inv}}(l, o_t, s_t, \hat{\mathbf{F}}_{t+n}). \tag{8}$$

The DiT head jointly models action distributions and progress values, capturing correlations among these cross-modal inputs across the trajectory. Training follows the standard diffusion objective:

$$\tilde{\mathbf{y}}_{t:t+n-1,t_d} = \sqrt{\bar{\alpha}_{t_d}}\, \mathbf{y}_{t:t+n-1} + \sqrt{1-\bar{\alpha}_{t_d}}\, \boldsymbol{\epsilon} \tag{9}$$

$$\mathcal{L}_{\text{DiT}} = \mathbb{E}_{t_d, \boldsymbol{\epsilon}} \left\| \boldsymbol{\epsilon} - \epsilon_\theta \left(\tilde{\mathbf{y}}_{t:t+n-1,t_d} \,\middle|\, l, o_t, s_t, \hat{\mathbf{F}}_{t+n}, t_d \right) \right\|_2^2 \tag{10}$$

where $\mathbf{y}_{t:t+n-1}$ is the target action–progress vector over steps t to $t+n-1$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is Gaussian noise, t_d is the diffusion time with $\bar{\alpha}_{t_d}$ the cumulative noise schedule, $\tilde{\mathbf{y}}_{t:t+n-1,t_d}$ is the noised target, and $\epsilon_\theta(\cdot)$ is the noise predictor.

4. Experiments

Training Details. Our training process consists of a pre-training and a fine-tuning stage. For pre-training, we utilize a mixed dataset from the DROID [54] and BridgeData V2 [113] datasets, which together provide large-scale, in-the-wild robotic arm demonstrations to build a foundational understanding of diverse real-world tasks, and EPIC-KITCHENS [20] and RoboCerebra [33], which provide fine-grained sub-steps and time-segment annotations to supervise the learning of semantic progress estimation in long-horizon, contact-rich scenarios. For fine-tuning, we select 942 trajectories from robot data and annotate them with affordance data and continuous progress labels using a semi-automated method. Details can be found in Appendix A.

4.1. Simulation Experiments

Benchmarks. We conduct evaluations across two simulation benchmarks: the **LIBERO** [77] benchmark compris-

Table 1. **CALVIN ABC→D experimental results.** We group the baselines into four types and report the average success rate of the top three checkpoints, computed over 1,000 rollouts per task, as well as the average number of consecutively completed tasks to solve 5 instructions (Avg. Len.). PALM consistently and substantially outperforms all baselines. Best results are shown in **bold**.

Method	Type	Task completed in a row					
		1	2	3	4	5	Avg. Len. ↑
RT-1 [10]	Autoregressive	53.3%	22.2%	9.40%	3.80%	1.30%	0.90
Robo-Flamingo [69]	Autoregressive	82.4%	61.9%	46.6%	33.1%	23.5%	2.47
OpenVLA [55]	Autoregressive	91.3%	77.8%	62.0%	52.1%	43.5%	3.27
Diffusion Policy [19]	Diffusion-based	40.2%	12.3%	2.60%	0.80%	0.00%	0.56
π_0 [8]	Diffusion-based	93.8%	85.0%	76.7%	68.1%	59.9%	3.92
3D-VLA [150]	3D-Aware	44.7%	16.3%	8.10%	1.60%	0.00%	0.71
3D Diffuser Actor [53]	3D-Aware	92.2%	78.7%	63.9%	51.2%	41.2%	3.27
RoboUniview [79]	3D-Aware	94.2%	84.2%	73.4%	62.2%	50.7%	3.65
Susie [7]	Prediction	87.0%	69.0%	49.0%	38.0%	26.0%	2.69
GR-1 [118]	Prediction	85.4%	71.2%	59.6%	49.7%	40.1%	3.06
Seer [112]	Prediction	94.4%	87.2%	79.9%	72.2%	64.3%	3.98
PALM (✗ progress)	Prediction	95.3%	85.6%	79.5%	74.3%	67.0%	4.02
PALM	**Prediction + Progress**	**96.9%**	**93.8%**	**89.3%**	**85.9%**	**82.0%**	**4.48**

ing four distinct task suites (Spatial, Object, Goal, Long), each with 10 tasks and 50 demonstrations for evaluating the robot's comprehension of spatial relationships, and the **CALVIN** [88] benchmark, designed to evaluate the instruction-following capabilities of robotic policies on long-horizon language-conditioned tasks, comprising 34 tasks across four distinct environments. We focus on the challenging ABC→D setting, where we pre-train in the ABC environments and evaluate in the unseen D environment.

Baselines. We compare PALM with four types of baselines: autoregressive methods, diffusion-based methods, 3D-aware methods and prediction methods. For the CALVIN ABC→D benchmark, we compare to RT-1 [10], Robo-Flamingo [69], and OpenVLA [55], which are autoregressive action models that generate actions from pre-trained VLMs. Diffusion Policy [19] and π_0 [8] are selected as representative diffusion-based methods that use a denoising diffusion process to model high-dimensional action distributions, while 3D-VLA [150], 3D Diffuser Actor [53], and RoboUniview [79] specialize in capturing 3D-aware representations to enhance manipulation. Prediction-based methods, represented by Susie [7], GR-1 [118], and Seer [112], merge visual foresight as a future representation to enhance performance in multi-task robot manipulation. For LIBERO, we compare with OpenVLA, Diffusion Policy, and Seer. Additionally, we compare against Octo [111], which pre-trains robot policies on diverse datasets to enhance generalization. SpatialVLA [100], CoT-VLA [148], TraceVLA [151], and CoA-VLA [67] are included to compare diverse task representation methods.

Results. Table 1 summarizes performance on the CALVIN ABC→D long-horizon benchmark, demonstrating that PALM achieves state-of-the-art results across all metrics and outperforms all baselines. First, PALM (Prediction + Progress) reaches a 96.9% success rate on the first subtask and maintains strong performance as horizon length increases (*e.g.*, 82.0% for five consecutive subtasks). This is a +17.7% absolute improvement over the strongest prior baseline (Seer at 64.3%) at the 5-task horizon. PALM also yields the longest average task trajectory (4.48), exceeding Seer (3.98) and π_0 (3.92), indicating significantly more stable execution over extended sequences. Importantly, results confirm that the progress-aware policy is critical for long-horizon generalization, as removing the progress prediction (PALM ✗ progress) reduces performance consistently across horizons (*e.g.*, average length 4.02 → 4.48).

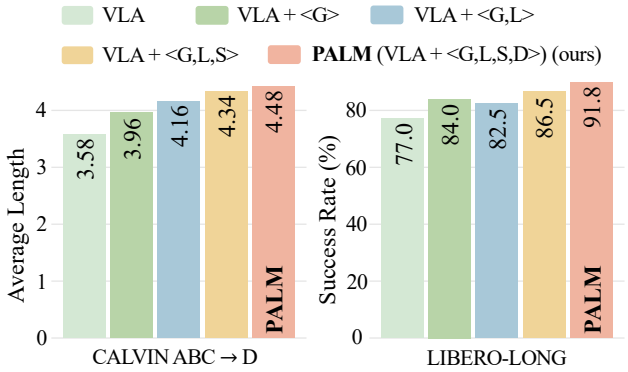


Figure 3. **Ablation studies of affordance components** on CALVIN ABC→D and LIBERO-LONG benchmarks demonstrate the effectiveness of the four components of affordance prediction.

Moreover, as shown in Table 2, across all four LIBERO suites, PALM achieves state-of-the-art performance with an average success rate of 94.5%. The largest gain is in LIBERO-LONG, where PALM reaches 91.8%, outperforming the strongest baseline (CoT-VLA at 69.0%) by 22.8%.

4.2. Ablation Studies

How do the components of the fine-grained affordance module affect performance? We evaluate the four

Table 2. **LIBERO experimental results.** For each task suite (Spatial, Object, Goal, Long), we report the average success rate and standard error across 3 seeds with 500 episodes each. PALM achieves the best performance over previous methods. Best results in **bold**.

	Average (↑)	Spatial (↑)	Object (↑)	Goal (↑)	Long (↑)
OpenVLA [55]	76.5 ± 0.6%	84.7 ± 0.9%	88.4 ± 0.8%	79.2 ± 1.0%	53.7 ± 1.3%
Diffusion Policy [19]	72.4 ± 0.7%	78.3 ± 1.1%	92.5 ± 0.7%	68.3 ± 1.2%	50.5 ± 1.3%
Octo fine-tuned [111]	75.1 ± 0.6%	78.9 ± 1.0%	85.7 ± 0.9%	84.6 ± 0.9%	51.1 ± 1.3%
SpatialVLA [100]	69.0 ± 1.2%	88.2 ± 0.7%	89.9 ± 1.3%	78.6 ± 0.9%	55.5 ± 1.5%
CoT-VLA [148]	81.1 ± 0.6 %	87.5 ± 1.4%	91.6 ± 0.5%	87.6 ± 0.6%	69.0 ± 0.8%
TraceVLA [151]	74.8 ± 0.9%	84.6 ± 1.0%	85.2 ± 0.6%	75.1 ± 1.4%	54.1 ± 1.0%
CoA-VLA [67]	79.8 ± 0.5%	85.3 ± 0.9%	93.1 ± 1.0%	85.8 ± 0.9%	55.0 ± 1.2%
PALM	**94.5 ± 1.0%**	**95.2 ± 1.2 %**	**96.7 ± 0.7%**	**94.3 ± 1.6%**	**91.8 ± 0.8%**

Table 3. **Ablation studies of PALM components.** Results on the CALVIN ABC→D benchmark demonstrate the effectiveness of each training module under both pre-training and fine-tuning.

Ablation Type	Pre-training	Fine-tuning
	Avg. Len. ↑	Avg. Len. ↑
PALM	**4.48**	**4.48**
✗ Affordance Foresight	3.90	3.58
✗ Inverse Dynamic Prediction	4.17	3.92
✗ Progress Prediction	3.73	4.02

Table 4. **Ablation studies on training data composition.** Results on the CALVIN ABC→D and LIBERO-LONG benchmarks demonstrate the data efficiency of each source type.

Ablation Type	CALVIN ABC→D	LIBERO-LONG
	Avg. Len. ↑	SR (%) ↑
PALM	**4.48**	**91.8**
✗ In-the-Wild Data	3.90	73.5
✗ Long-Horizon Video Data	3.73	84.5
✗ Human Annotated Data	3.58	76.5
✗ Simulation Data (Pretrain)	3.96	81.0

types of predictable affordance latents: `Global`, `Local`, `Spatial`, and `Dynamic` to measure their respective contributions. We train models by cumulatively adding these affordances to a vanilla VLA. As illustrated in Figure 3, adding Global affordance (G) already yields a consistent improvement on both CALVIN and LIBERO-LONG, indicating that coarse object-centric cues help stabilize long-horizon reasoning. Incorporating Local affordance (L) provides additional gains on CALVIN but introduces a slight dip on LIBERO-LONG, likely due to viewpoint-induced geometric bias that affects fine-grained edge-based features. Adding Spatial affordance (S) restores and further improves performance across both benchmarks by providing robust placement priors that generalize across layout variations. Finally, adding Dynamic affordance (D) to the full set (PALM) yields the best performance, indicating that combining motion cues with structured spatial reasoning produces the most reliable long-horizon behavior.

How do the proposed PALM modules affect performance? We study the contributions of PALM's three core components (affordance prediction module, inverse dynamics prediction, and progress-aware prediction) by conducting ablations during both pre-training and fine-tuning on the CALVIN ABC→D benchmark. Table 3 shows that all three PALM components contribute complementary capabilities and collectively result in the best performance. Removing affordance foresight produces the largest drop in fine-tuning performance (4.48 → 3.58), confirming that structured future affordance prediction is essential for precise long-horizon planning once the policy is adapted to downstream robot data. In contrast, removing progress prediction causes the largest degradation during pre-training (4.48 → 3.73), highlighting that large-scale long-horizon datasets are particularly valuable for learning a strong progress prior that stabilizes temporal reasoning. Eliminating inverse dynamics prediction also reduces performance in both stages, showing that predicting multi-step action trajectories conditioned on affordance latents provides an important training signal beyond direct behavior cloning.

Effectiveness of training data composition for robot manipulation. Training data is a critical factor influencing the performance of robot policies. Accordingly, we classify our training data into four primary types based on their task orientation: (1) In-the-wild datasets, represented by DROID [54] and BridgeData V2 [113] datasets for the pre-training stage, (2) Long-horizon video datasets, consisting of EPIC-KITCHENS [20] and RoboCerebra [33], which are also utilized during pre-training to learn the understanding of semantic progress in complex, multi-step scenarios, (3) Human-annotated datasets, employed in the fine-tuning stage to train affordance and progress-aware policy, and (4) Simulation datasets for post-training evaluation (CALVIN ABC→D [88] and LIBERO-LONG [77]).

As illustrated in Table 4, excluding any single data type invariably degrades model performance, though to varying extents. This degradation is most pronounced for the In-the-Wild and Human-Annotated datasets. A substantial performance decrease also occurs when removing the pre-training data corresponding to the simulation benchmarks. Additional ablation studies are provided in Appendix C.

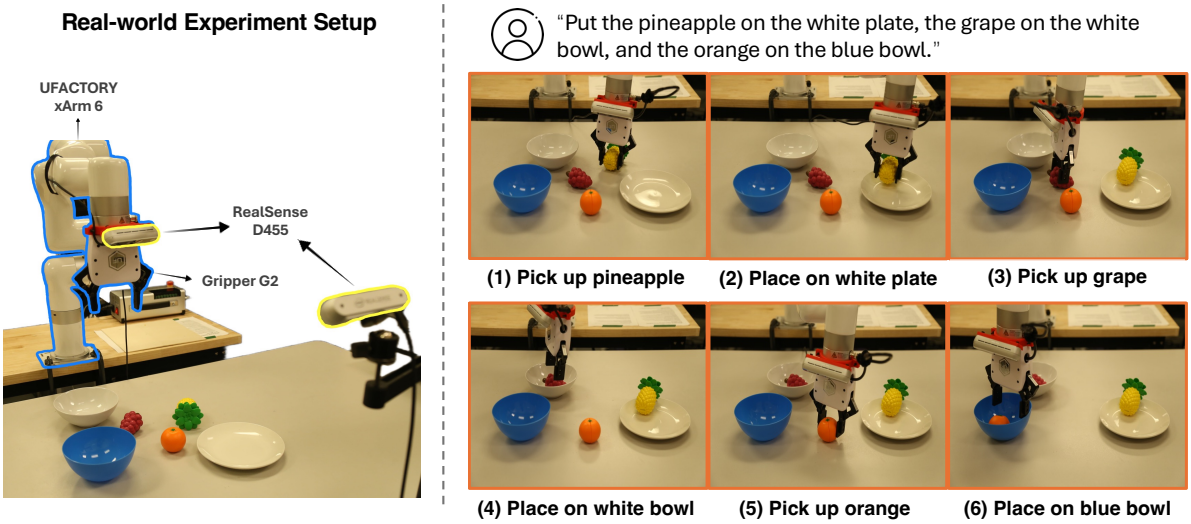


Figure 4. **Real-world experimental setup and task design. Left:** We use a UFACTORY xArm6 robot with the matched Gripper G2 and two RealSense D455 cameras. **Right:** We design a real-world long-horizon manipulation task consisting of six consecutive subtasks, driven by a single high-level instruction.

Table 5. **Real-world experimental results** on long-horizon task under different generalization settings.

Type	Method	Task completed in a row						Avg. Len. ↑
		1	2	3	4	5	6	
Random Localization	OpenVLA [55]	0.45	0.30	0.15	0.05	0.00	0.00	0.95
	Octo [111]	0.35	0.20	0.10	0.00	0.00	0.00	0.65
	PALM	**0.70**	**0.65**	**0.55**	**0.45**	**0.40**	**0.30**	**3.05**
Visual Distraction	OpenVLA [55]	0.65	0.50	0.25	0.15	0.05	0.00	1.60
	Octo [111]	0.45	0.35	0.15	0.00	0.00	0.00	0.95
	PALM	**0.85**	**0.80**	**0.65**	**0.60**	**0.50**	**0.40**	**3.80**
Unseen Lighting	OpenVLA [55]	0.55	0.35	0.25	0.10	0.00	0.00	1.25
	Octo [111]	0.50	0.35	0.15	0.05	0.00	0.00	1.05
	PALM	**0.80**	**0.70**	**0.60**	**0.60**	**0.45**	**0.40**	**3.55**

4.3. Real-World Experiments

Experimental Setup. As shown in Figure 4, we use a UFACTORY xArm6 with a Gripper G2 to conduct real-world experiments. For visual input, we use two RealSense D455 cameras configured as eye-on-hand and eye-on-base to capture RGB images. We design several types of generalization tests for long-horizon tasks, including Visual Distraction, Random Localization, and Unseen Lighting. We select a mixed pre-training dataset composed of DROID [54] and BridgeData V2 [113], while the fine-tuning dataset consists of 200 demonstrations collected on the xArm with RGB images, robot states, and actions.

Baselines and Metrics. We select OpenVLA [55] and Octo [111] as baselines. We report the success rate (SR) and average length for each task over 20 real-world rollouts. For each rollout, each method is permitted a maximum of three execution attempts. To ensure fairness, all models are fine-tuned on our training dataset, trained for an equal number of iterations, and evaluated with the final checkpoint.

Generalization Evaluation for Long-Horizon Task. To evaluate generalization in long-horizon tasks, we construct a sequential pick-and-place task consisting of 6 consecutive subtasks, driven by a single high-level instruction. We consider three generalization settings: ① varying the target object pose, ② changing the scene lighting to unseen conditions, and ③ adding multiple distractor objects to induce visual clutter. As shown in Table 5, results demonstrate PALM's superior generalization over baselines as the task sequence length increases, showing its robustness in long-horizon settings. Details are available in Appendix B.

5. Conclusion

We introduce **PALM**, a vision–language–action (VLA) model for long-horizon robotic manipulation that couples structured future affordance prediction with continuous progress estimation in a closed loop. PALM achieves state-of-the-art results on two benchmarks, with a 12.5% improvement on CALVIN ABC→D and 91.8% success on LIBERO-LONG, and shows significant robustness in real-world experiments across long-horizon generalization settings.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 2

[2] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13778–13790, 2023. 3

[3] Zechen Bai, Chen Gao, and Mike Zheng Shou. Evolve-vla: Test-time training from environment feedback for vision-language-action models. *arXiv preprint arXiv:2512.14666*, 2025. 3

[4] Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language. *arXiv preprint arXiv:2403.01823*, 2024. 2

[5] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024. 2

[6] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. 2

[7] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023. 2, 6

[8] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 2, 6

[9] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[10] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 2, 6

[11] Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. Goal-conditioned reinforcement learning with imagined subgoals. In *International Conference on Machine Learning (ICML)*, pages 1430–1440, 2021. 3

[12] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024. 2

[13] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465, 2024. 5

[14] Changan Chen, Juze Zhang, Shrinidhi K Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The language of motion: Unifying verbal and non-verbal language of 3d human motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6200–6211, 2025. 7

[15] Joya Chen, Difei Gao, Kevin Qinghong Lin, and Mike Zheng Shou. Affordance grounding from demonstration video to target image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6799–6808, 2023. 3

[16] Qianzhong Chen, Justin Yu, Mac Schwager, Pieter Abbeel, Yide Shentu, and Philipp Wu. Sarm: Stage-aware reward modeling for long horizon robot manipulation. *arXiv preprint arXiv:2509.25358*, 2025. 2

[17] Xinyue Chen, Zijian Zhou, Zheng Wang, Che Wang, Yanqiu Wu, and Keith Ross. Bail: Best-action imitation learning for batch deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18353–18363, 2020. 3

[18] Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, et al. Villa-x: enhancing latent action modeling in vision-language-action models. *arXiv preprint arXiv:2507.23682*, 2025. 2, 3

[19] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025. 2, 6, 7

[20] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 43 (11):4125–4141, 2020. 1, 2, 5, 7

[21] Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Wenhao Zhang, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*, 2025. 2

[22] Zhihao Dou, Qinjian Zhao, Zhongwei Wan, Dinggen Zhang, Weida Wang, Towsif Raiyan, Benteng Chen, Qingtao Pan, Yang Ouyang, Zhiqiang Gao, et al. Plan then action: High-level planning guidance reinforcement learning for llm reasoning. *arXiv preprint arXiv:2510.01833*, 2025. 2

[23] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 2

[24] Songcheng Du, Yang Zou, Zixu Wang, Xingyuan Li, Ying Li, Changjing Shang, and Qiang Shen. Unsupervised hyperspectral image super-resolution via self-supervised modality decoupling. *International Journal of Computer Vision*, 134 (4):152, 2026. 7

[25] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023. 2

[26] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9156–9172, 2023. 2

[27] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021. 2

[28] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024. 2

[29] Caelan Reed Garrett, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Pddlstream: Integrating symbolic planners and blackbox samplers via optimistic adaptive planning. In *Proceedings of the International Conference on Automated Planning and Scheduling*, pages 440–448, 2020. 3

[30] Yiran Geng, Boshi An, Haoran Geng, Yuanpei Chen, Yaodong Yang, and Hao Dong. End-to-end affordance learning for robotic manipulation. *arXiv preprint arXiv:2209.12941*, 2022. 3

[31] Jiayuan Gu, Sean Kirmani, Paul Wohlhart, Yao Lu, Montserrat Gonzalez Arenas, Kanishka Rao, Wenhao Yu, Chuyuan Fu, Keerthana Gopalakrishnan, Zhuo Xu, et al. Rt-trajectory: Robotic task generalization via hindsight trajectory sketches. *arXiv preprint arXiv:2311.01977*, 2023. 2

[32] Qiao Gu, Yuanliang Ju, Shengxiang Sun, Igor Gilitschenski, Haruki Nishimura, Masha Itkina, and Florian Shkurti. Safe: Multitask failure detection for vision-language-action models. *arXiv preprint arXiv:2506.09937*, 2025. 7

[33] Songhao Han, Boxiang Qiu, Yue Liao, Siyuan Huang, Chen Gao, Shuicheng Yan, and Si Liu. Robocerebra: A large-scale benchmark for long-horizon robotic manipulation evaluation. *arXiv preprint arXiv:2506.06677*, 2025. 1, 2, 5, 7

[34] Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang, Yanbiao Ma, Yunfeng Diao, Ziyu Jia, Wenbo Ding, Hangjun Ye, and Long Chen. Roboafford++: A generative ai-enhanced dataset for multimodal affordance learning in robotic manipulation and navigation. *arXiv preprint arXiv:2511.12436*, 2025. 3

[35] Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, Guang Li, Zheng Lu, Shuhuai Ren, Xianhui Meng, et al. Mimo-embodied: X-embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*, 2025. 2

[36] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 3, 1

[37] Zhi Hou, Tianyi Zhang, Yuwen Xiong, Haonan Duan, Hengjun Pu, Ronglei Tong, Chengyang Zhao, Xizhou Zhu, Yu Qiao, Jifeng Dai, et al. Dita: Scaling diffusion transformer for generalist vision-language-action policy. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7686–7697, 2025. 2

[38] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024. 2, 3

[39] Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 9488–9495. IEEE, 2024. 2

[40] Siyuan Huang, Haonan Chang, Yuhan Liu, Yimeng Zhu, Hao Dong, Peng Gao, Abdeslam Boularias, and Hongsheng Li. A3vlm: Actionable articulation-aware vision language model. *arXiv preprint arXiv:2406.07549*, 2024. 3

[41] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022. 2

[42] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024. 2

[43] Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. Pointworld: Scaling 3d world models for in-the-wild robotic manipulation. *arXiv preprint arXiv:2601.03782*, 2026. 7

[44] Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachy Groom, Hunter Hancock, Karol Hausman, Gashon Hussein, Brian Ichter, Szymon Jakubczak, Rowan Jen, Tim Jones, Ben Katz, Liyiming Ke, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Yao Lu, Vishnu Mano, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Charvi Sharma, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Alex Swerdlow, James Tanner, Marcel Torne, Quan Vuong, Anna Walling, Haohuan Wang, Blake Williams, Sukwon Yoo, Lili Yu, Ury Zhilinsky, and Zhiyuan Zhou. $\pi^*_{0.6}$: a vla that learns from experience. *arXiv preprint arXiv:2511.14759*, 2025. 7

[45] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. 2

[46] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning (ICML)*, pages 4651–4664, 2021. 3, 1

[47] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1734, 2025. 2

[48] Yu-qian Jiang, Shi-qi Zhang, Piyush Khandelwal, and Peter Stone. Task planning in robotics: an empirical comparison of pddl-and asp-based systems. *Frontiers of Information Technology & Electronic Engineering*, 20(3):363–373, 2019. 3

[49] Yuanchen Ju, Kaizhe Hu, Guowei Zhang, Gu Zhang, Mingrun Jiang, and Huazhe Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *European Conference on Computer Vision (ECCV)*, pages 222–239. Springer, 2024. 3

[50] Xuhui Kang and Yen-Ling Kuo. Incorporating task progress knowledge for subgoal generation in robotic manipulation through image edits. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7490–7499. IEEE, 2025. 2

[51] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European Conference on Computer Vision (ECCV)*, pages 18–35, 2024. 5

[52] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023. 2

[53] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*, 2024. 2, 6

[54] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024. 1, 2, 5, 7, 8

[55] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 2, 6, 7, 8, 5

[56] Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026. 7

[57] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 4

[58] Yuxuan Kuang, Junjie Ye, Haoran Geng, Jiageng Mao, Congyue Deng, Leonidas Guibas, He Wang, and Yue Wang. Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation. *arXiv preprint arXiv:2407.04689*, 2024. 2, 3

[59] Yuzhi Lai, Shenghai Yuan, Peizheng Li, Jun Lou, and Andreas Zell. Seer-var: Semantic egocentric environment reasoner for vehicle augmented reality. *arXiv preprint arXiv:2508.17255*, 2025. 7

[60] Bingyu Li, Haocheng Dong, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Exploring efficient open-vocabulary segmentation in the remote sensing. *arXiv preprint arXiv:2509.12040*, 2025.

[61] Bingyu Li, Feiyu Wang, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Maris: Marine open-vocabulary instance segmentation with geometric enhancement and semantic alignment. *arXiv preprint arXiv:2510.15398*, 2025.

[62] Bingyu Li, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation. In *ACM International Conference on Multimedia (ACM MM)*, pages 1308–1317, 2025. 7

[63] Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026. 2

[64] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10922–10931, 2023. 3

[65] Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3086–3096, 2024. 3

[66] Gen Li, Nikolaos Tsagkas, Jifei Song, Ruaridh Mon-Williams, Sethu Vijayakumar, Kun Shao, and Laura Sevilla-Lara. Learning precise affordances from egocentric videos for robotic manipulation. In *International Conference on Computer Vision (ICCV)*, pages 10581–10591, 2025. 3

[67] Jinming Li, Yichen Zhu, Zhibin Tang, Junjie Wen, Minjie Zhu, Xiaoyu Liu, Chengmeng Li, Ran Cheng, Yaxin Peng, Yan Peng, et al. Coa-vla: Improving vision-language-action

models via visual-text chain-of-affordance. In *International Conference on Computer Vision (ICCV)*, pages 9759–9769, 2025. 2, 3, 6, 7

[68] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026. 7

[69] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023. 2, 6

[70] Xiang Li, Cristina Mata, Jongwoo Park, Kumara Kahatapitiya, Yoo Sung Jang, Jinghuan Shang, Kanchana Ranasinghe, Ryan Burgert, Mu Cai, Yong Jae Lee, et al. Llara: Supercharging robot learning data for vision-language policy. *arXiv preprint arXiv:2406.20095*, 2024. 2

[71] Zhaoyang Li, Zhan Ling, Yuchen Zhou, Litian Gong, Erdem Bıyık, and Hao Su. Oric: Benchmarking object recognition under contextual incongruity in large vision-language models. *arXiv preprint arXiv:2509.15695*, 2025. 2

[72] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[73] Min Lin, Xiwen Liang, Bingqian Lin, Liu Jingzhi, Zijian Jiao, Kehan Li, Yuhan Ma, Yuecheng Liu, Shen Zhao, Yuzheng Zhuang, et al. Echovla: Robotic vision-language-action model with synergistic declarative memory for mobile manipulation. *arXiv preprint arXiv:2511.18112*, 2025. 3

[74] Tao Lin, Gen Li, Yilei Zhong, Yanwen Zou, Yuxin Du, Jiting Liu, Encheng Gu, and Bo Zhao. Evo-0: Vision-language-action model with implicit spatial understanding. *arXiv preprint arXiv:2507.00416*, 2025. 2

[75] Tao Lin, Yilei Zhong, Yuxin Du, Jingjing Zhang, Jiting Liu, Yinxinyu Chen, Encheng Gu, Ziyan Liu, Hongyi Cai, Yanwen Zou, et al. Evo-1: Lightweight vision-language-action model with preserved semantic alignment. *arXiv preprint arXiv:2511.04555*, 2025. 2

[76] Suhan Ling, Yian Wang, Ruihai Wu, Shiguang Wu, Yuzheng Zhuang, Tianyi Xu, Yu Li, Chang Liu, and Hao Dong. Articulated object manipulation with coarse-to-fine affordance for mitigating the effect of point cloud noise. In *International Conference on Robotics and Automation (ICRA)*, pages 10895–10901. IEEE, 2024. 3

[77] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 44776–44791, 2023. 1, 2, 5, 7

[78] Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-world robotic manipulation through mark-based visual prompting. *arXiv preprint arXiv:2403.03174*, 2024. 3

[79] Fanfan Liu, Feng Yan, Liming Zheng, Chengjian Feng, Yiyang Huang, and Lin Ma. Robouniview: Visual-language model with unified view representation for robotic manipulation. *arXiv preprint arXiv:2406.18977*, 2024. 6

[80] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024. 2

[81] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pages 38–55, 2024. 4

[82] Weifeng Lu, Minghao Ye, Zewei Ye, Ruihan Tao, Shuo Yang, and Bo Zhao. Robofac: A comprehensive framework for robotic failure analysis and correction. *arXiv preprint arXiv:2505.12224*, 2025. 7

[83] Teli Ma, Zifan Wang, Jiaming Zhou, Mengmeng Wang, and Junwei Liang. Glover: Generalizable open-vocabulary affordance reasoning for task-oriented grasping. *arXiv preprint arXiv:2411.12286*, 2024. 3

[84] Teli Ma, Jia Zheng, Zifan Wang, Ziyao Gao, Jiaming Zhou, and Junwei Liang. Glover++: Unleashing the potential of affordance learning from human behaviors for robotic manipulation. *arXiv preprint arXiv:2505.11865*, 2025. 3, 4

[85] Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. Liv: Language-image representations and rewards for robotic control. In *International Conference on Machine Learning (ICML)*, pages 23301–23320, 2023. 3

[86] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 655–677, 2023. 3

[87] Lucas Manuelli, Wei Gao, Peter Florence, and Russ Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. In *The International Symposium of Robotics Research*, pages 132–157. Springer, 2019. 3

[88] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3): 7327–7334, 2022. 1, 2, 6, 7

[89] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *International Conference on Computer Vision (ICCV)*, pages 6813–6823, 2021. 3

[90] Tongzhou Mu, Zhaoyang Li, Stanisław Wiktor Strzelecki, Xiu Yuan, Yunchao Yao, Litian Liang, and Hao Su. When should we prefer state-to-visual dagger over visual reinforcement learning? In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 14637–14645, 2025. 2

[91] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022. 3

[92] Soroush Nasiriany, Sean Kirmani, Tianli Ding, Laura Smith, Yuke Zhu, Danny Driess, Dorsa Sadigh, and Ted Xiao. Rt-

affordance: Affordances are versatile intermediate representations for robot manipulation. In *International Conference on Robotics and Automation (ICRA)*, 2025. 2, 3

[93] Dantong Niu, Yuvan Sharma, Giscard Biamby, Jerome Quenum, Yutong Bai, Baifeng Shi, Trevor Darrell, and Roei Herzig. Llarva: Vision-action instruction tuning enhances robot learning. *arXiv preprint arXiv:2406.11815*, 2024. 2

[94] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024. 2

[95] Qingtao Pan, Zhihao Dou, and Shuo Li. Frequency-modulated visual restoration for matryoshka large multimodal models. *arXiv preprint arXiv:2603.11220*, 2026. 7

[96] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4195–4205, 2023. 3, 1

[97] Karl Pertsch, Oleh Rybkin, Jingyun Yang, Shenghao Zhou, Konstantinos Derpanis, Kostas Daniilidis, Joseph Lim, and Andrew Jaegle. Keyframing the future: Keyframe discovery for visual prediction and planning. In *Learning for Dynamics and Control*, pages 969–979. PMLR, 2020. 3

[98] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7587–7597, 2024. 3

[99] Zengyi Qin, Kuan Fang, Yuke Zhu, Li Fei-Fei, and Silvio Savarese. Keto: Learning keypoint representations for tool manipulation. In *International Conference on Robotics and Automation (ICRA)*, pages 7278–7285. IEEE, 2020. 3

[100] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025. 6, 7

[101] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3, 1

[102] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 3, 1

[103] Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. *arXiv preprint arXiv:2407.05996*, 2024. 2

[104] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *arXiv preprint arXiv:2506.21656*, 2025. 2

[105] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, pages 894–906. PMLR, 2022. 2

[106] Noah Y Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, Nicolas Heess, and Martin Riedmiller. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*, 2020. 3

[107] Priya Sundaresan, Suneel Belkhale, Dorsa Sadigh, and Jeannette Bohg. Kite: Keypoint-conditioned policies for semantic manipulation. *arXiv preprint arXiv:2306.16605*, 2023. 2

[108] Huajie Tan, Sixiang Chen, Yijie Xu, Zixiao Wang, Yuheng Ji, Cheng Chi, Yaoxu Lyu, Zhongxia Zhao, Xiansheng Chen, Peterson Co, et al. Robo-dopamine: General process reward modeling for high-precision robotic manipulation. *arXiv preprint arXiv:2512.23703*, 2025. 3

[109] Yingbo Tang, Lingfeng Zhang, Shuyi Zhang, Yinuo Zhao, and Xiaoshuai Hao. Roboafford: A dataset and benchmark for enhancing object and spatial affordance learning in robot manipulation. In *ACM International Conference on Multimedia (ACM MM)*, pages 12706–12713, 2025. 3

[110] Yingbo Tang, Shuaike Zhang, Xiaoshuai Hao, Pengwei Wang, Jianlong Wu, Zhongyuan Wang, and Shanghang Zhang. Affordgrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9433–9439. IEEE, 2025. 3

[111] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. 2, 6, 7, 8, 5

[112] Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. *arXiv preprint arXiv:2412.15109*, 2024. 2, 3, 6

[113] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, pages 1723–1736, 2023. 1, 2, 5, 7, 8

[114] Qiang Wang, Robert McCarthy, David Cordova Bulens, Kevin McGuinness, Noel E O'Connor, Francisco Roldan Sanchez, Nico Gürtler, Felix Widmaier, and Stephen J Redmond. Improving behavioural cloning with positive unlabeled learning. In *Conference on Robot Learning (CoRL)*, pages 3851–3869, 2023. 3

[115] Yongxin Wang, Meng Cao, Haokun Lin, Mingfei Han, Liang Ma, Jin Jiang, Yuhao Cheng, and Xiaodan Liang. Eaco: Enhancing alignment in multimodal llms via critical observation. *arXiv preprint arXiv:2412.04903*, 2024. 7

[116] Yongxin Wang, Zhicheng Yang, Meng Cao, Mingfei Han, Haokun Lin, Yingying Zhu, Xiaojun Chang, and Xiaodan

Liang. Care what fails: Contrastive anchored-reflection for verifiable multimodal. *arXiv preprint arXiv:2512.19554*, 2025. 7

[117] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023. 2

[118] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023. 2, 6

[119] Pengyuan Wu, Pingrui Zhang, Zhigang Wang, Dong Wang, Bin Zhao, and Xuelong Li. Closed-loop action chunks with dynamic corrections for training-free diffusion policy. *arXiv preprint arXiv:2603.01953*, 2026. 3

[120] Yueh-Hua Wu, Nontawat Charoenphakdee, Han Bao, Voot Tangkaratt, and Masashi Sugiyama. Imitation learning from imperfect demonstration. In *International Conference on Machine Learning (ICML)*, pages 6818–6827, 2019. 3

[121] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022. 2

[122] Wenli Xiao, Haotian Lin, Andy Peng, Haoru Xue, Tairan He, Yuqi Xie, Fengyuan Hu, Jimmy Wu, Zhengyi Luo, Linxi Fan, et al. Self-improving vision-language-action models with data generation via residual rl. *arXiv preprint arXiv:2511.00091*, 2025. 7

[123] Feng Xu, Guangyao Zhai, Xin Kong, Tingzhong Fu, Daniel FN Gordon, Xueli An, and Benjamin Busam. Stare-vla: Progressive stage-aware reinforcement for fine-tuning vision-language-action models. *arXiv preprint arXiv:2512.05107*, 2025. 2

[124] Haoran Xu, Xianyuan Zhan, Honglei Yin, and Huiling Qin. Discriminator-weighted offline imitation learning from suboptimal demonstrations. In *International Conference on Machine Learning (ICML)*, pages 24725–24742, 2022. 3

[125] Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, Quan Vuong, Chelsea Finn, Dorsa Sadigh, and Sergey Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. *arXiv preprint arXiv:2402.19432*, 2024. 2

[126] Yuyin Yang, Zetao Cai, Yang Tian, Jia Zeng, and Jiangmiao Pang. Gripper keypose and object pointflow as interfaces for bimanual robotic manipulation. *arXiv preprint arXiv:2504.17784*, 2025. 2

[127] Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. Llm4drive: A survey of large language models for autonomous driving. *arXiv preprint arXiv:2311.01043*, 2023. 7

[128] Zhenjie Yang, Yilin Chai, Xiaosong Jia, Qifeng Li, Yuqian Shao, Xuekai Zhu, Haisheng Su, and Junchi Yan. Drivemoe: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving. *arXiv preprint arXiv:2505.16278*, 2025.

[129] Zhenjie Yang, Xiaosong Jia, Qifeng Li, Xue Yang, Maoqing Yao, and Junchi Yan. Raw2drive: Reinforcement learning with aligned world models for end-to-end autonomous driving (in carla v2). *arXiv preprint arXiv:2505.16394*, 2025. 7

[130] Lei Yao, Yong Chen, Yuejiao Su, Yi Wang, Moyun Liu, and Lap-Pui Chau. Hammer: Harnessing mllm via cross-modal integration for intention-driven 3d affordance grounding. *arXiv preprint arXiv:2603.02329*, 2026. 3

[131] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026. 7

[132] Jianjian Yin, Tao Chen, Gensheng Pei, Huafeng Liu, Yazhou Yao, Liqiang Nie, and Xiansheng Hua. Semi-supervised semantic segmentation with multi-constraint consistency learning. *IEEE Transactions on Multimedia*, 2025. 7

[133] Tianjiao Yu, Xinzhuo Li, Yifan Shen, Yuanzhe Liu, and Ismini Lourentzou. Core3d: Collaborative reasoning as a foundation for 3d intelligence. *arXiv preprint arXiv:2512.12768*, 2025. 7

[134] Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026. 7

[135] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024. 2, 5

[136] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024. 2

[137] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024. 2

[138] Jia Zeng, Qingwen Bu, Bangjun Wang, Wenke Xia, Li Chen, Hao Dong, Haoming Song, Dong Wang, Di Hu, Ping Luo, et al. Learning manipulation by predicting interaction. *arXiv preprint arXiv:2406.00439*, 2024. 2

[139] Shaopeng Zhai, Qi Zhang, Tianyi Zhang, Fuxian Huang, Haoran Zhang, Ming Zhou, Shengzhe Zhang, Litao Liu, Sixu Lin, and Jiangmiao Pang. A vision-language-action-critic model for robotic real-world reinforcement learning. *arXiv preprint arXiv:2509.15937*, 2025. 3

[140] Hongyin Zhang, Zifeng Zhuang, Han Zhao, Pengxiang Ding, Hongchao Lu, and Donglin Wang. Reinbot: Amplifying robot visual-language manipulation with reinforcement learning. *arXiv preprint arXiv:2505.07395*, 2025. 7

[141] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 516–526, 2024. 7

[142] Juze Zhang, Changan Chen, Xin Chen, Heng Yu, Tiange Xiang, Ali Sartaz Khan, Shrinidhi K Lakshmikanth, and Ehsan Adeli. Vibes: A conversational agent with behaviorally-

intelligent 3d virtual body. *arXiv preprint arXiv:2512.14234*, 2025. 7

[143] Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiang Zhu, and Jianyu Chen. Up-vla: A unified understanding and prediction model for embodied agent. *arXiv preprint arXiv:2501.18867*, 2025. 2

[144] Kaifeng Zhang, Zhao-Heng Yin, Weirui Ye, and Yang Gao. Learning manipulation skills through robot chain-of-thought with sparse failure guidance. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 8012–8018. IEEE, 2025. 2

[145] Pingrui Zhang, Yifei Su, Pengyuan Wu, Dong An, Li Zhang, Zhigang Wang, Dong Wang, Yan Ding, Bin Zhao, and Xuelong Li. Cross from left to right brain: Adaptive text dreamer for vision-and-language navigation. *arXiv preprint arXiv:2505.20897*, 2025. 7

[146] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, Fan Lu, He Wang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *arXiv preprint arXiv:2507.04447*, 2025. 2

[147] Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 729–741, 2024. 7

[148] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1702–1713, 2025. 2, 6, 7

[149] Wei Zhao, Pengxiang Ding, Min Zhang, Zhefei Gong, Shuanghao Bai, Han Zhao, and Donglin Wang. Vlas: Vision-language-action model with speech instructions for customized robot manipulation. *arXiv preprint arXiv:2502.13508*, 2025. 2

[150] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024. 2, 6

[151] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024. 3, 6, 7

[152] Pengfei Zhou, Liliang Chen, Shengcong Chen, Di Chen, Wenzhi Zhao, Rongjun Jin, Guanghui Ren, and Jianlan Luo. Act2goal: From world model to general goal-conditioned policy. *arXiv preprint arXiv:2512.23541*, 2025. 3

[153] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025. 2

[154] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 2

[155] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, pages 2165–2183, 2023. 2